DOCUMENT RESUME

ED 307 244 SP 031 149

AUTHOR Handley, Herbert M., Ed.
TITLE Evaluation: Test Construction and Use. An Instructional Model for Undergraduate Teacher Education in the RAFT Program at Mississippi State University.
INSTITUTION Mississippi State Univ., Mississippi State. Bureau of Educational Research and Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Dec 86
CONTRACT 400-85-1053
NOTE 52p.; For related documents, see SP 031 144-148.
PUB TYPE Reports - Descriptive (141) -- Guides - Classroom Use - Materials (For Learner) (051)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Academic Achievement; Educational Objectives; Higher Education; Preservice Teacher Education; Rating Scales; *Standardized Tests; Student Behavior; *Student Evaluation; *Teacher Made Tests; *Test Construction
IDENTIFIERS Mississippi State University; *Research Applications for Teaching Project

ABSTRACT

This module developed by the Research Applications for Teaching (RAFT) project assists the preservice teacher in constructing test items to better measure the outcomes of instructional objectives. Student teachers are also assisted in the interpretation of results of a student's performance on a standardized test. Students also "trouble-shoot" a real test used by a teacher in their field to evaluate achievement. Careful attention is given to using alternative methods of evaluation of student behaviors in the affective and psychomotor domains. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

EVALUATION: TEST CONSTRUCTION AND USE

An Instructional Model Prepared for Undergraduate

Teacher Education in the RAFT Project

at Mississippi State University

by

Herbert M. Handley, Editor

Department of Curriculum and Instruction

David T. Morse, Contributor

Department of Educational Psychology

Mississippi State University

This book was developed under Contract No. 400-85-1053, funded by the National Institute of Education, monitored and published by the Bureau of Educational Research and Evaluation at Mississippi State University. Points of view or opinions contained herein do not necessarily represent official position or policy of Mississippi State University or the National Institute of Education.

December 1986

TABLE OF CONTENTS

SECTION	PAGE
HOW TO USE THIS MODULE	ii
LESSON 1: EVALUATION PROCESSES	1
Objectives	1
A Definition	1
Seeking Appropriate Information-Gathering Instruments	3
Types of Data Gathering Instruments	5
Advantages and Disadvantages of Test Items	7
WRITING GOOD TEST ITEMS	9
Selected Response Items	11
Multiple Choice Items	12
True-False Items	19
Matching Items	22
Other Item Types	23
General Rules for Writing Selected Response Items	23
Constructed Response Items	24
Short Answer Items	26
Essay Items	28
General Rules for Writing Constructed Response Items	30
General Rules for Item Writing	30
Other Data about Tests	33
ASSIGNMENT 1: Constructing Test Items	35
ASSIGNMENT 2: Evaluating a Test's Structure	36
ASSIGNMENT 3: Constructing a Rating Scale	37

SECTION PAGE

LESSON 2: INTERPRETING STANDARDIZED TEST RESULTS 38

Comprehensive Tests of Basic Skills (CTBS). 39

Measurement Terms and Concepts. 39

Types of Scores 40

Limitations/Cautions in Interpreting Test Scores. 41

Student Diagnostic Profile. 42

TEST ON TEST CONSTRUCTION 43

BIBLIOGRAPHY. 46

HOW TO USE THIS MODULE

This module of instruction was planned to assist teachers to write better developed test items to measure the outcomes of instructional objectives. Students are also assisted in the interpretation of results regarding a student's performance on a standardized test. Students also "trouble-shoot" a real test used by a teacher in their field to evaluate achievement of students in a small unit of instruction.

Students are also encouraged to work especially hard on the assignments in this module since the acquisition of well-developed test construction skills is a prerequisite for teachers in today's world of criterion-referenced testing.

Lesson 1

EVALUATION PROCESS

Objectives

The student will:

1. Define evaluation and describe each of the four stages in the evaluation process.
2. Demonstrate cognition of the appropriate information-gathering instruments when seeking to make classroom evaluations.
3. Write good test items for evaluating achievement.
4. Develop checklists and rating scales for evaluating student products and performance.
5. Describe how to use information to grade, to judge student progress, to judge changes in student attitudes, and to judge the effectiveness of the instructional program.

A Definition

To evaluate is to place a value upon--to judge. However, forming judgement is not an isolated action. Information is needed before informed judgement is made. Furthermore, making an informed judgement is necessary prior to making a decision. To put it another way, evaluation is the process of obtaining information and using it to form judgements which, in turn, are used in decision making.

Preparing for evaluation. In preparing for evaluation of student progress, the teacher decides the kind of information needed. A determination is made concerning when and how to obtain

the needed information. The instructional objectives suggest the type of information needed. Evaluative schema used must parallel the instructional objectives advanced.

Obtaining needed information. A wide variety of information is gathered to evaluate students' progress. This information may be obtained from students' responses to criterion-related tests, standardized achievement tests, observation checklists, or through observations of other aspects of students' classroom performance. Collection of data relative to affective behaviors may also be appropriate.

Forming judgements. After analyzing all information obtained, judgements are made by comparing the information to selected criteria reflecting expectations of each student's performance. Such judgements may be concerned with whether or not each of the respective students is performing above, on or below grade level, with primary weaknesses in the student's classroom performance, with the cause of an individual's learning problems or with the attitudes of an individual toward his or her work.

Use judgements to make decisions and evaluation reports. In this area of evaluation, the teacher records significant findings and plans an appropriate course of action for future education of students. Findings regarding specific students are filed appropriately in the school's records. Findings relative to a specific student are shared with his or her parents. Their cooperation is sought, and hopefully, secured.

Seeking Appropriate Information-Gathering Instruments

After determining what will be evaluated and what information is needed in order to evaluate students, the teacher is ready to choose an instrument for obtaining that information.

There are basically four different techniques classroom teachers use to obtain information about themselves and their students: inquiry, observation, analysis and testing. To inquire is to ask. Good teachers are always asking students how they feel about what is going on. Through inquiry several types of information are secured, such as opinions, self-perceptions of students, subjective judgements, affective behaviors and social perceptions. Information secured by inquiry is the least objective kind of data available and is highly subject to human bias and error. Collecting data by inquiry in the classroom is an inexpensive process monetarily, but securing it may be costly in terms of time consumed.

Observations of students' performance in the classroom are done routinely by teachers. When the results of observations are recorded in a systematic way, these data are very useful in evaluation of pupil progress. Observations provide data relative to the performance of students or the end products of some performance. Emotional behaviors are best evaluated through use of observations. Observations also provide data for evaluating progress in areas where formal testing is difficult, such as in early childhood education where formal testing is difficult to execute. Observations of students' performance in music, art,

shops, physical educations and science laboratories also provide valuable data. Such observations also include studies of behaviors crossing more than one domain of learning (cognitive, affective, psychomotor). Observational data do have characteristics of being somewhat subjective, but these data can be made more objective by careful planning of observation instruments. Doing observations is very time consuming but relatively inexpensive.

Teachers also secure more data about students by analyzing the products of their performance. This process may include, for example, the analysis of a piece of woodwork prepared in a class in vocational education, an examination of a written piece of work to discover inconsistencies in sentence structure or a breakdown of a mathematical problem's solution to determine the type of errors being made. Such analyses may be used to determine the type of errors being made. Such analyses may be used to determine learning outcomes at both intermediate (during the learning process) and summative stages. Data secured from such analyses are objective, but the results may not be stable over time; i.e., if the data were analyzed again at a later time by the teacher, a different interpretation might be obtained.

Testing is the most frequently used procedure for obtaining data relative to students' academic progress. Testing is used whenever there is a common situation to which all students respond (e.g., a test question), a common set of instructions structuring the students' responses, a set of rules for correctly scoring the responses, and a criteria reporting each student's performance (a

score). Testing provides data relative to the attitudes and achievement of students. It provides data for assessing terminal goals, maximum performance of students and cognitive outcomes in general. Data derived from testing is the most objective and reliable evaluative material available. Testing provides more information per unit of time than any other evaluation techniques, but this data gathering process is the most expensive one available.

Types of Data Gathering Instruments

Five widely used types of instruments used for data collection are: (1) standardized tests; (2) teacher-made tests; (3) checklists; (4) rating scales; and (5) questionnaires.

Standardized tests are used when very accurate information is needed. Students respond to standardized tests in very similar conditions. Most of these tests are commercially available in a highly competitive market. Hence, they usually have been carefully developed and field tested. Data on reliability and validity are available. Reliability of the test refers to accuracy of the test's data and to whether or not the test measures what it purports to measure. Normative data are usually available to interpret a student's achievement on the test relative to a national, regional or state sample. A disadvantage in the use of standardized tests for evaluation of students' progress is that these tests often do not measure exactly what has been taught in a local setting. The achievement tested comes from a broader spectra of cognitive performance. Use of standardized tests is an

expensive process and use of data derived is limited to what is measured by the test.

Teacher-made tests are used routinely to obtain achievement information. These tests, in their best form, are criterion-referenced; i.e., they measure exactly what has been taught. They are inexpensive and can be constructed with relative facility. Disadvantages in using teacher-made tests are that no norms are available beyond the class tested. Also, teacher-made tests may take a long time to construct; and, unless the teacher is skilled in test-making, these tests are often unreliable.

Checklists are used to structure observations. They are helpful in organizing observations around key points or critical behaviors of interest. Checklists, however, can measure only the presence or absence of an observed trait or behavior. Examples of when checklists would be useful would be listing the criteria of a good speech and checking off how well the student performs according to these criteria on his speech, listing the qualities of a good science project and checking off how many of these criteria a student demonstrated in his work or listing the types of process skills young children are expected to demonstrate (observing, classifying, hypothesizing; describing, inferring) and then checking off those they are observed doing satisfactorily.

Rating scales are used to judge the quality of a performance. They are useful in making quality judgements, as well as quantitative judgements, about students' performances. Rating scales can be used to assess the quality of a speech given, the

quality of a metal piece constructed in shop, the quality of a painting or the quality of a planned constitution for a group's operation. Good and valid rating scales take a great deal of time and effort to construct. They also can be clumsy to interpret, if they are not carefully constructed.

Questionnaires are used to inquire about feelings, opinions and interests of students. They are advantageous in that they keep inquiry focused and help the teacher obtain the same type of information about each student. Unfortunately, they take time and effort to construct. They are difficult to score since there are no right or wrong answers. Hence, data are difficult to summarize.

Advantages and Disadvantages of Test Items

Five common types of test items are written: (1) short answer; (2)essay; (3) true/false; (4) matching and (5) multiple choice.

Short answer. Short answer problems may call for solution to problems in mathematics, labeling the parts of a flower, listing the basic five food groups, filling in the correct terms in blanks. These types of test items can measure achievement at both low and high order thinking skill areas. Short answer items can test many facts in a short time and are fairly easy to score. Short answer items are an excellent format for mathematics. These items generally test recall. On the other hand, it is difficult to measure complex learning with short answer questions. These items, also, are ambiguous.

Essay items. Essay items can test complex learning. They can

be used to evaluate thinking processes and creativity. Essay questions are difficult to score objectively. Responses to these items require a long testing time. Also, essay questions require more time for scoring.

True/False items. True/false items can test more facts in a short time. These tests are objective and easy to score. True/false items test recognition. It is difficult to measure complex learning with true/false items. It is also difficult to write reliable items. Responses to true/false items are subject to guessing.

Matching. Matching items are excellent for testing associations and recognition of facts. Although terse, these items can test complex learning, especially concepts. These items are objective in form. It is very difficult to write good matching items. Also, if the items are not properly written, responses are influenced by process of elimination.

Multiple choice. These items can evaluate learning at all levels of complexity. Multiple choice items can be highly reliable and objective. Fairly large knowledge bases can be tested in a short time with these items. Multiple choice tests are easy to score. Multiple choice items are difficult to write and responses are somewhat subject to guessing.

WRITING GOOD TEST ITEMS

by

David T. Morse

Why should teachers even worry about writing items when so many commercially-prepared tests are available? One reason is that available commercially-published tests typically do not measure the specific objectives of a particular course, so locally-made tests must be used. Most teachers, not having taken a course in classroom test construction, must rely upon intuition and personal experience for guidance in test construction. However, there are many poor test writing practices in use which can result in inaccurate information about the student's capabilities, so that both the student and instructor are shortchanged. Writing good test items is achieved by: (a) understanding the advantages and disadvantages of the different item types; (b) considering a few basic principles of item writing; and (c) taking the time to carefully plan, construct and review the test items.

Writing good test items is as easy as falling off a log. However, those readers who have fallen off logs know just how painful that can be. (This probably explains why professional test makers have such glum expressions.) The remainder of this section will discuss the different item types, good practices to follow for each type, and some general rules for constructing good test items and tests. With these suggestions as a guide, perhaps item-writing needn't be as distressing a task.

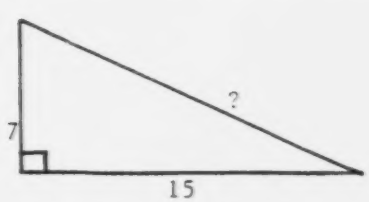
There are two major types of test items: the selected response and the constructed response type. The item types differ in terms of what kind of response the student must give. Many persons believe, incorrectly, that selected response items (i.e., multiple-choice, true-false, or matching) can only assess memorization skills, while constructed response items (i.e., short answer or essay) are more appropriate for measuring so-called "higher order skills." Often this is referred to as "recall vs. recognition," where the constructed response items are considered to be measuring recall, and the selected response items are considered to measure only recognition. In point of fact, most cognitive skills can be measured by a variety of test items. Consider the following examples.

Type of Item

1. What is the capital of Oregon?

_____

2.



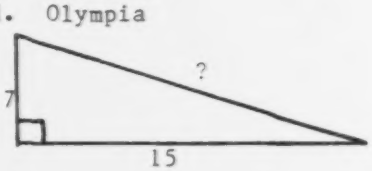
What is the length of the hypotenuse in the right triangle above?

_____

1. What is the capital of Oregon?

a. Eugene
b. Salem
c. Portland
d. Olympia

2.



What is the length of the hypotenuse in the right triangle above?

a. 13.6
b. 16.6
c. 22.0
d. 22.6

Both items numbered one measure recall of a simple fact--that the capital of Oregon is Salem. Both items numbered two measure the application of the Pythagorean theorem ($c = a + b$). The recall-recognition argument holds only for knowledge-level skills and not for higher-order skills. True, there is a chance that students could guess the correct answer to the second selected-response item, and this topic will be discussed later.

Selected Response Items

Selected response items include multiple-choice, true-false, matching, and other types of items in which the student is expected to select or order the correct responses. There are several advantages of selected response items. These include:

1. Ease and replicability of scoring. Selected response items can be marked by almost anyone given the answer key. Further, the high degree of objectivity in the scoring reduces the effects of subtle biases in the scoring process.

2. Student response rate. Students can respond to a greater number of selected response items than constructed response items in a given period of time. Thus, a selected response test can cover more material than can a constructed response test.

3. Measurement of desired skills only. Many times, teachers give constructed response items even though they do not intend to rate the student's ability to construct good paragraphs, spell correctly, or produce clear explanations. When this occurs, the teacher has made the task more difficult and time-consuming--by

having added the constructed response burden--than it would be in a selected response format. The rationale for posing such items in a constructed response format is often questionable.

4. Adaptability. Should a selected response item prove to be poorly written or otherwise unfair to the students, it can be thrown out of the test with much less loss of information than could an extended essay item, for instance. Also, selected response tests are much more adaptable to the use of machine-scorable answer sheets.

The primary disadvantages of selected response items are:

1. Subject to guessing. Students are sometimes able to correctly guess on selected response items. In fact, if a student guesses at random, the expected score on a true-false test, for instance, is 50%. The cure for this problem is to include more items, which means the chance of guessing one's way to a passing mark is reduced, or for some items, requiring that the students show their work.

2. More time-consuming to write. Selected response tests are more time-consuming to write than are constructed response tests or items requiring the same amount of time to answer.

Now that the primary advantages and disadvantages of selected response items have been discussed, we will examine specific types of selected response items.

Multiple-Choice Items

Multiple-choice items present more than one alternative

response from which the student is to select an answer. The true-false item is actually a special type of multiple-choice item having only two alternatives. Likewise, the matching item is a special type of multiple-choice item in which the same alternatives are used. The multiple-choice item is perhaps the best-known type of selected response item. There are several rules to follow when writing multiple-choice items.

1. Present the problem in a clear and unambiguous form. Compare these two examples.

Poor	Good
Plants live because of:	The biochemical process by which plants sustain life is called:

The second item stem is much less ambiguous--without even looking at the alternatives, the student immediately has a better idea as to what kind of answer will be correct. The "stem" of a multiple-choice item refers to the lead statement(s).

2. Avoid the use of specific determiners. A specific determiner is a characteristic of a poorly-written item which tends to give away the answer. Compare the following examples. (Note: Correct answers are denoted by an asterisk.)

Poor	Good
How have scientists recognized the great work of Linnaeus?	How have scientists recognized the great work of Linnaeus?
a. by giving him the Nobel prize b. by founding a college with his name	a. by awarding him the Nobel prize for his work

*c. by adding the letter L. to the names of all the animals he had classified
d. by awarding him a cash prize

b. by founding a college of natural science with his name
*c. by adding L. to the scientific name of animals he classified
d. by establishing a cash prize to be awarded for outstanding achievement

The first example is artificially easy because the correct response is so much longer and contains more information than do the other alternatives. The second example requires a little more reading, but now the alternatives are much closer to one another in terms of length and amount of information. Likewise, using a very short correct answer relative to other alternatives should be avoided.[1]

Poor	Good
Another word for convivial is:	Another word for convivial is:
a. voracious	a. trivial
b. inextricable	b. inextricable
c. placebo	c. vitiate
*d. jovial	*d. sociable

In the first example, the word convivial in the item stem bears some resemblance to the correct response, jovial, both in the spelling and sound of the last two syllables. This is called an alliterative association. An uninformed student could guess the correct answer on the basis of this similarity. By changing the word jovial to sociable, and changing alternatives (a) and (c), the

[1]There is research which indicates that even elementary school-age children are capable of detecting item construction flaws such as this and others described in this manual. See Morse, 1980.

second example avoids this problem. An uninformed student using the same strategy for this item would likely choose the wrong answer.

Consider a more suitable type of specific determiner, illustrated by the following examples:

Poor	Good
An ichthyologist is a person who:	An ichthyologist is a person who studies:
*a. studies fish.	*a. fish.
b. plays in mud puddles.	b. ants.
c. sells real estate.	c. industrial pollution.
d. makes chemical compounds.	d. rock formations.

The first example could be answered correctly by the uninformed student who recognizes that the article a in the item stem can match only to an alternative beginning with a consonant. Also, the incorrect alternatives, while rarely in the vocabulary of most folks, are too far removed from the subject. In the second example, a simple change removes the grammar cue, and the new alternatives have some relationship to methods of travel.

These are the three most common types of specific determiners: length of alternative; alliterative association; and grammatical cues. These flaws should be avoided.

3. Avoid the use of"hang-on" alternatives. There are very few occasions in which alternatives such as "all of these" and "none of these" are required. Compare the following examples.

Poor	Good
Which of the following is (are) characteristic of anerobic bacteria?	Which of the following is a characteristic of anerobic bacteria?
a. manufactures chlorophyll *b. lives without oxygen c. reproduces sexually d. all of the above e. none of the above f. a and c only	a. manufactures chlorophyll *b. lives without oxygen c. reproduces sexually

There is no compelling argument for trying to include an equal number of response alternatives for all multiple-choice items. An item-writer should not struggle to produce, say, four alternatives for each item. If three plausible choices are all that can be created, then do not waste your time trying to make up another. The first example above serves to waste the student's time by requiring much more reading than is necessary. More often than not, the use of "all of the above" or "none of the above" is not required.[1]

4. Avoid redundant reading in the items. Often the amount of reading required by the student as well as the space taken up by an item can be reduced. Compare these two examples.

[1]One might wonder why such alternatives are so popular in commercially published tests. According to one test publisher, the reason is so that the item will more closely approximate an "infinite-choice" (i.e., constructed response) item. An alternative explanation, say, for mathematics items might be to avoid the case of a student being able to choose an answer which seems "reasonable" rather than working the solution. Force of tradition or the need to have a uniform member of response alternatives both seem equally likely explanations of the practice.

The Reconstruction era in U.S. history:	During what term as U.S. President did the Reconstruction era come to an end?
(a) ended during the term of Abraham Lincoln as President	a. Abraham Lincoln
(b) ended during the term of William H. Taft as President	b. William H. Taft
(c) ended during the term of Theodore Roosevelt as President	c. Theodore Roosevelt
*(d) ended during the term of Rutherford B. Hayes as President	*d. Rutherford B. Hayes

Both examples pose the very same problem, but the second item is much shorter and easier to read.

5. Keep numerical alternatives in a logical order. Items which have only numerical alternatives are easier to read and respond to if the alternatives are in a logical order. Compare the following examples.

Poor	Good
3/8 : 1 1/2 = ?	3/8 : 1 1/2 = ?
a. 1 1/8	*a. 1/4
b. 9/16	b. 9/16
c. 15/16	c. 15/16
*d. 1/4	d. 1 1/8
e. 1 7/8	e. 1 7/8

The second item requires less reading time from the student. Notice that the item alternatives could also have been placed in descending order and still be less confusing to follow than those

in the first item.

6. Avoid including some common error in logic in an item. Consider the following items.

Poor	Good
In order to be a U. S. senator, a person must be at least (in years):	What is the minimum age (in years) as set by the Constitution for a person to be a U. S. senator?
a. 21	a. 21
b. 25	b. 25
*c. 30	*c. 30
d. 35	d. 35
e. 40	e. 40

Note that the first example is worded in such a way that any of the first three alternatives can be considered correct. That is, a person must be at least 21 years of age in order to be a U. S. senator. A slight change in the stem, in the second example, clears up the error in logic and makes the answer unarguable.

7. Avoid giving alternatives which have the same meaning, and are therefore incorrect. Compare the following examples.

Poor	Good
When the temperature drips below 32° F:	Which of the following events typically occurs when the temperature is below 32° F?
a. water will freeze	a. Ice melting
b. ice will form	*b. Water freezing
c. snow will sometimes fall	c. Tornado formation
*d. all of the above	d. Liquid precipitation

In the first example, alternative (a) and (b) mean essentially the same thing. Thus, the correct answer to the item must be some other choice. Because alternative (c) is also true, the answer has to be (d). This problem has been corrected in the second example

by having included only one correct alternative. A second issue pertinent to the examples above is that of the "best answer." Rather than selecting a correct response, the student must choose the best answer in the first example. Selection of best answer is a different task altogether than is selection of a correct answer.

8. Avoid including silly or nonsense alternatives for an item. An illustration of such a practice is given in the following examples.

Poor	Good
Which of these people has been Governor of Mississippi?	Which of these people has been Governor of Mississippi?
a. Jimmy Carter	a. Evelyn Gandy
b. Queen Elizabeth	b. C. B. Newman
c. Erik Estrada	c. John Stennis
*d. Theodore Bilbo	*d. Theodore Bilbo

The first example includes some absurd alternatives which most students could reject outright because they know better. The second example, while not perfect--the incorrect choices are recent office-holders--at least provides more plausible alternatives. If your goal is to be the wry test-item writer, then you may wish to consider an occasional "off-the-wall" alternative. However, other than the short-term comic relief (whether real or imagined), all such alternatives tend to do is take up extra space and reading time.

True-False Items

The true-false item can usually be answered even more quickly than a multiple-choice item. There are several rules to follow for

constructing true-false items.

1. Be sure the item is absolutely true or false. The student who is aware of exceptions to a particular statement will be confused as to which way to respond to the item.

2. Avoid the use of specific determiners. Many students are aware of the typically false response called for by any statement containing the word always or never. That is, such "absolute" words often tip off the test-wise student. To include such items serves little purpose. By the same token, "qualified" words such as often, may be, seldom, many and few can tip off the otherwise uninformed student that the corresponding statement will generally be correct. However, it is sometimes a good practice to include such items when the correct answer is contrary to the general pattern. Consider the following examples.

Poor	Good
All birds can fly.	All birds have a type of wing structure.

The uninformed student could very likely determine that the first example was an incorrect statement. The second example is a correct statement which incorporates an absolute word. As such, it is contrary to the general pattern. The uninformed student is likely to be misled.

3. Avoid the use of negatives or double negatives. These tend to make the item more difficult than it would otherwise be. Compare the following examples.

Poor	Good
It is incorrect to suggest that the theme of Macbeth is not concerned with the human vices of greed and ambition.	The theme of Macbeth deals with the human vices of greed and ambition.

By the time the student correctly deciphered the first item, probably half a dozen items like the second one could have been answered.

4. Avoid inclusion of "double-barreled" items. A double-barreled item is one which calls for two distinct judgements. The following examples provide an illustration.

Poor	Good
Because nearly all birds fly south for the winter, the migration patterns for most birds are well-known.	The migratory patterns of most species of birds in the Northern Hemisphere have been mapped.

The first example poses a double-barreled item. The phrase, "Because nearly all . . ." can be judged as incorrect by the student. But now a dilemma: Is the teacher's intent for the student to judge the accuracy of the clause "the migration patterns . . . ," the introductory phrase, or both? The second example, in addition to clearing up some content-related quibbles, has limited itself to a single proposition to be judged by the student.

5. Include a larger number of false than true items. Students who guess on true-false items tend to select true more often than they select false. This fact unfairly favors the uninformed student who is given a test primarily composed of true items.

Matching Items

A matching item is actually a condensed set of multiple-choice items, each having the same set of response alternatives from which to choose. There are four basic rules to follow for matching items.

1. Make the directions for the item clear and complete. Compare these two examples.

Poor	Good
Match the following:	Write the letter of each answer listed on the right in the blank by the proper statement on the left. Some answers may be used more than once.

The second set of directions more clearly sets the task for the student.

2. Use more alternatives than there are items to match or allow answers to be used more than once. This prevents the students from obtaining correct answers by the process of elimination.

3. Hold matching sections to a maximum of ten or twelve statements (premises). Matching sections having more than a dozen items require too much time of the student for scanning the alternatives.

4. Construct matching sections such that the individual items are related. For example, in a social studies test, avoid making a matching section which includes persons, dates, places, treaties and battles. Instead, make short, separate matching sections, one

on persons, one on dates, and so forth.

5. Arrange at least one column in alphabetical order; this saves time in searching for the correct answer.

Other Item Types

There are other types of items which can be classified as selected response items. In nearly all cases, the general rules of providing clear directions and posing the question in an unambiguous form are the most important to follow. This leads us to some general rules for selected response items.

General Rules for Writing Selected Response Items

The goal of all well-written test items should be to measure what the student knows, not what give-away items he or she can detect, or any other test-taking ability unrelated to the content. Towards this end, there are some general rules applicable for all types of selected response items.

1. Always provide clear directions to the student as to how to respond to the items.

2. Keep the vocabulary level of the items as simple as possible. Unless the test is meant to measure reading comprehension, there is little need to use large words when smaller ones will suffice.

3. Write the items so that they present the task or pose the problem to the student in a clear (unambiguous, not transparent) manner.

4. Keep the items independent of each other. Avoid writing a series of items in which the answer to one item determines how the following item will be answered.

5. Write items covering important concepts and not trivial information. One sure way to construct a test measuring the acquisition of trivial information is to write the items at the last minute without careful thought and planning.

6. Avoid trick questions. Unless you are trying to teach students how to cope with trick questions or are trying to measure I.Q., including them can only serve to confuse the students.

7. Avoid writing items which give the answer to other items in the same test. It should be obvious that such items are not a fair measure of what the student knows.

8. The answers to selected response items should be those to which other content experts would agree. If no consensus can be obtained, the item could most likely stand to be revised.

9. If you can't answer the question, don't expect the students to be able to answer it.

Constructed Response Items

Constructed response items include short answer, completion, short essay and extended essay items, in which the expected student behavior is construction of an appropriate response. There are several advantages of constructed response items as a class. These include:

1. Reduced guessing. The likelihood of successful guessing

on items by students is much smaller than for selected response items.

2. Good measures of writing ability. Constructed response tests are good for assessing the student's ability to construct clear sentences, write paragraphs, spell vocabulary words, and so on.

3. Easy to prepare. Constructed response items are faster and easier to write than are selected response items calling for the same behaviors.

The primary disadvantages of constructed response items are:

1. More time consuming to score. Constructed response tests require more time to score by the teacher. The use of machine-scorable forms is not feasible.

2. More time consuming to complete. Typically, a given number of constructed response items will require more time for a student to complete than would the same number of selected response items.

3. Sensitive to scoring bias. Many studies have shown that ratings on constructed response items are affected by such extraneous factors as handwriting ability, spelling accuracy, and even by such different raters for constructed response than for selected response items. For this reason, constructed response tests are sometimes referred to as "subjective tests."

4. Amount of topic coverage. The amount of material which can be covered in a given length of time using constructed response items is less than for selected response items.

Now that the primary advantages and disadvantages of constructed response items have been discussed, we will examine specific types of constructed response items.

Short Answer Items

Short answer items require a response ranging from a few words to a complete sentence. The completion item is one type of short answer item which calls for one or more responses to be inserted in order to complete a sentence or phrase. There are several rules to follow in the preparation of short answer items.

1. Include directions for the student to follow. Should the student prepare a complete sentence, or will key words suffice? Should the student write in the blanks in a question or underneath the item? These are simple considerations, but important for informing the student as to what response is expected.

2. State the item as precisely as possible. This is an especially critical point for completion ("fill-in-the-blank") items. Compare the following examples.

Poor	Good
How is a woodpecker like a frog? _____	What are two biological needs which woodpeckers and frogs have in common? _____

A student might believe that any number of answers to the first item could be justified, such as both eat insects or both are animals. The second example is much less likely to be misinterpreted.

Poor	Good
Modern devices make _____ more effective.	What are three scientific devices which have helped astronomers in their work?

The first example is ambiguous and virtually any human endeavor, such as tennis, could be argued as a good answer. The second much more clearly defines the task. It is often recommended that the completion items be constructed so that the "blank" is at the end of the statement, and not in the middle or at the beginning. Remember--the purpose of sound test items is to determine what the student has learned, not how well the student can decipher confusing test items.

3. Try to avoid writing completion items requiring multiple responses. Compare the following examples.

Poor	Good
A _____ telescope has a concave _____ and an eyepiece _____.	What type of telescope has a concave mirror and an eyepiece lens? _____

The first item is so thoroughly mutilated as to be unintelligible. It also bears a suspicious resemblance to a sentence lifted verbatim out of a textbook with a few words replaced by blanks. While such items would be one way to measure whether students have memorized their reading assignment, they are not necessarily measuring anything beyond that sort of mindless rote learning. The second example is much more clear. It is very difficult indeed to write good completion items which call for multiple inserts from the student; the best practice is to avoid

such items.

4. Prepare answer keys for the items. When preparing answer keys for short answer items, try to include all the possible responses which could be considered correct. This will prevent a lengthy scoring period, since you will not have to mull over new responses trying to decide if they are acceptable or not.

Essay Items

Essay items require a response ranging from a few sentences, for short essay items, to several paragraphs for extended essay items. There are several rules to follow in the preparation of essay items.

1. Include directions for the student to follow. Some instructors prefer essay answers in complete sentence form, while others prefer outline form. Be sure to make the directions explicit enough that the student will be aware of what type and length of response is required.

2. Include guidelines for responding and scoring. It is often helpful to fully explain how a student's response will be evaluated. Will spelling count? Will points be added or subtracted for overall appearance? How many specific points or examples are to be included in the answer? Compare the examples which follow.

Poor	Good
List the results of World War II. Use both sides of this sheet if necessary.	In 300 words or less, describe two political and two economic results of World War II. Be sure to give a specific example for each result. A complete answer is worth 10 points. 2 1/2 points will be deducted for each missing or incorrect result. Use complete sentences. Spelling will not count in your mark.

The first example is so incomplete that any student would be hard-pressed to produce an acceptable response. The second example defines the task much more completely.

3. Avoid using essay items for simple listing tasks. Writing a good essay requires a considerable amount of effort and time on the part of the student. That time and effort would be better spent on items requiring a thoughtful response than on items which call for a recitation of knowledge-level information. Compare the following examples.

Poor	Good
List the confederate states and give the dates of secession.	In one paragraph, explain the major reasons why the confederate states chose to secede. A complete answer will include two reasons and is worth 5 points. Use complete sentences.

The first example, measuring whether students have memorized a large number of states and dates, would be better posed in a short answer format. The second example presents a task much better suited to the essay format.

4. Prepare a model answer sheet for use in scoring. Answer

the item yourself, being sure to include all the specific facts, examples, and so on, which make up an acceptable response. Then compare your answer to the question. They should match in terms of length, number of examples, and other requirements. Also, having the model answer is helpful for the actual rating of student papers. By making the testing and scoring process more objective, some of the hidden sources of bias associated with essay items can be reduced.

General Rules for Writing Constructed Response Items

The goal of well-written constructed response items should be the presentation of a clear, unambiguous task to the student. Such items will have a greater likelihood of measuring the intended outcomes of the instruction. To this end, there are some general rules applicable to all types of constructed response items.

1. Always provide clear directions to the student as to how to respond to the items.

2. Keep the vocabulary level of the items as simple as possible.

3. Write the items so that they present the task or pose the problem in a clear and unambiguous manner.

4. Prepare sample answers to check both the clarity of the item and use as a model for grading student papers.

General Rules for Item Writing

There are some general guidelines to follow in the

construction of any type of test item.

1. Have others review your items. Allowing other instructors to look at your items is a quick and convenient method for verifying both the clarity of the item and the accuracy of the keyed or sample response.

2. Be prepared to revise items. Few test items are as good as they can possibly be the first time they are written. If knowledgeable students have difficulty with particular items, the items might well be faulty in some respect.

3. Include clear directions to the student.

4. Avoid making questions from verbatim quotations from a text or other source. More often than not, such items are poor, and tend to measure only retention of trivial information.

5. Avoid trick questions. Trick questions tend to measure skills unrelated to the content.

6. Make the items in the test reflect the relative importance and time spent on each topic. This is one of the purposes of the test specifications. If the vast majority of the course was spent covering the use of left-handed widgets, then the vast majority of items on the test should also cover the use of left-handed widgets.

7. Avoid negatives and double negatives in test items. Many times, a simple rewording of such items makes them easier to comprehend.

In summary, a well-written test should be easy--that is, answerable--for the informed student. To the extent that test items are artificially easy or difficult for the student, both the

student and instructor are shortchanged. Following relatively simple guidelines such as those presented here is inexpensive insurance against writing confusing or misleading test items.

Some of you will encounter objectives which are not measurable using paper and pencil tests. Instead, a performance test is necessary. A brief summary of how to create performance tests is given in Appendix E.

OTHER DATA ABOUT TESTS

Measures of Central Tendency

In describing results of test data for groups, mean, median and model test scores are often reported. These data are useful for teachers to know in order to interpret the results of the total group's performance.

Mean. The mean score for the group is a simple average score. The sum of scores for the entire group is totaled and divided by the number of people in the group. The mean or average score of a group on a test is probably most representative of the performance for the group as a whole.

Median. The median score is the score lying at the very center of the distribution of the group's scores. Exactly the same number of individuals scored higher than the median as did those who scored lower. This score is interpreted as the middle performance level of the group.

Mode. The modal score is that which was earned by more of the students in the group than any other score. It is the most "popular" score.

Upper and Lower Quarters

Students are often described as placing in the upper quarter, middle half, or lower quarter on a test. Those placing in the upper quarter earned a score which placed them above the 75th percentile according to the norm group. This group is usually

considered the higher achievers and are often offered enrichment activities. Those in the lower quarter scored at the 25th percentile or lower. They are the "at risk" group. Remedial activities are planned for students falling in this group. The middle half of the students range in scores from the 26th percentile to the 75th percentile. These students are usually considered as scoring in the average range on the test.

Assignment 1: Constructing Test Items

Directions: Please construct the following described test items according to procedures outlined in the preceding written material developed by Morse.

1. Construct two essay items, one of which requires at least a two-page response and the other requiring a discussion of one paragraph.
2. Construct four multiple-choice test items, one of which tests cognitive learnings at each of the following levels: knowledge, comprehension, application and analysis.
3. Construct two sets of matching test items.
4. Construct five true-false items.
5. Construct five selected response items.

Assignment 2: Evaluating a Test's Structure

Directions: Given a test constructed by a teacher in your subject area, study it carefully and evaluate it relative to the following criteria. To complete this assignment, you will also need a copy of the teacher's instructional objectives.

1. Do the questions correctly assess cognitive learnings at the level elicited in the teacher's objectives?

2. Are the items constructed according to the rules outlined by Morse?

3. Does the test evaluate learnings at cognitive levels more advanced than the basic knowledge and comprehension levels?

4. Is the test neatly constructed and is the length appropriate for grade and subject level for which the test is planned?

Assignment 3: Constructing a Rating Scale

Directions: Select a rating scale or check-list to be used in evaluating specific student behaviors which are learning outcomes in either the affective or the psychomotor domain. This instrument, for example, might be a rating scale for evaluating the quality of a science project, an oral presentation made by a student in class or a physical performance in physical education. Construct an instrument that would be useful in your own teaching area.

Lesson 2

INTERPRETING STANDARDIZED TEST RESULTS

Objectives

The student will:

1. Interpret the meaning of results from a student's taking a standardized test.

2. Tell the difference in a norm-referenced test and a criterion-referenced test.

3. Differentiate between the mean, mode and median in test scores.

4. Explain the meaning of the following terms:
   A. Upper quarter; lower quarter
   B. Stanine score
   C. Normal curve equivalent
   D. Grade equivalent score
   E. Percentile rank on national norms

REPORTING, INTERPRETING AND APPLYING TEST RESULTS

Information from standardized tests can be useful in the process of curricular and instructional planning to benefit children. The results from a test that measures student achievement should provide indicators of acquired skills, to date, in two formats:

* Norm-Referenced Information - identifies general strengths and weaknesses of groups or individuals by subtest relative to a national comparison group,

* Objectives Mastery Information - identifies more specific strengths and weaknesses of groups or individuals relative to instructional objectives measured.

Measurement Terms and Concepts

To use test scores and reports effectively, it is important to have a basic understanding of test and measurement terms. The descriptions that follow are not intended to be technically detailed but, rather, are meant to provide a short, general reference source for use with CTBS results.

Norm-Referenced Tests

A norm-referenced test provides information derived from a predetermined group, called the norm group, whose characteristics are known and described. Scores of a particular group are compared with scores of the norm group. Norm-referenced information is obtained by converting scale scores to the derived scores of interest. CTBS/U is a norm-referenced achievement test that also provides criterion-referenced information.

Criterion-Referenced Tests

A criterion-referenced test provides information on individual or group mastery of objectives reflecting specific skills. Mastery scores reflect what the student knows or can do rather than how the student compares to a reference group. CTBS/U is a norm-referenced achievement test that also provides criterion-referenced information.

Achievement Tests

An achievement test is a test designed to identify the knowledge and skills that students have acquired in specified content areas at a certain point in time. Achievement tests can be norm-referenced or criterion-referenced, or they can include elements of both, as does CTBS/U.

Note--This material is included in the module with permission of publishers of the Comprehensive Test of Basal Skills.

Types of Scores

Scale Scores (SS)

The scale score is the basic score for CTBS. It is used primarily to provide a basis for deriving other normative scores to describe test performance.

Scale scores are units of a single, equal-interval scale that is applied across all levels of CTBS, regardless of grade or time of year of testing. These scores are expressed in numbers that can range from 0 through 999. The equal-interval property of the scale makes these scores especially appropriate for various statistical purposes.

The principal limitation of scale scores is that they are not well suited to direct interpretation of individual performance. Therefore, the primary use of CTBS scale scores is to provide a basis for deriving the various other scores that can be used to describe test performance.

Percentile Ranks (NP or LP)

Percentile ranks, which range from 1 to 99, are commonly used for reporting test results to students and parents. A percentile rank may be interpreted as a percentage of students in a norm group whose scores fall below a given student's scale score. For example, if a student's scale score converts to a percentile rank of 71, this may be interpreted to mean that the student scored higher than approximately 71 per cent of the students in the norm group. Local percentiles may also be computed based on the distribution of scores in the local student population.

Stanines (NS)

Stanines are standard scores based on a scale of nine equal units that range from a high of 9 to a low of 1. In general, stanines of 1 through 3 are considered below average, 4 through 6 average, and 7 through 9 above average.

A stanine is less precise than a percentile rank, but it is relatively easy to work with and to interpret.

Objective Mastery score (OMS)

The objective mastery score, OMS, is a criterion-referenced score that reflects a student's mastery of the test objectives. The OMS is reported as a ratio of the number of objectives mastered to the total number of objectives for the content area. For example, an objective mastery score for Reading Comprehension might be reported as 4/7, indicating that the student mastered four out of seven objectives in the Reading Comprehension subtest.

Grade Equivalent (GE)

A grade equivalent, GE, is a score expressed in terms of grade and month. It indicates the grade and month in school of students in the norm group whose performance most nearly typifies that of a given student. For example, if a second grader obtained a grade equivalent of 4.8 on a mathematics subtest, that grade equivalent would not mean that the student had mastered all the mathematics that is taught in the school during the first eight months of Grade 4. It would mean only that the student's performance on that test was theoretically equivalent to the typical performance of students in the norm group who had completed eight months of Grade 4.

Since grade equivalents do not indicate ability level, they are not appropriate for use in placing students in school grades or instructional programs.

Limitations/Cautions in Interpreting Test Scores

It is important to remember that when scores are being analyzed and interpreted, the results are descriptions of an individual's or group's performance at a single point in time (See Standard Error of Measurement, TCS description). Scores can fluctuate upon repeated testing, and should therefore be interpreted with a range of possible scores in mind, rather than an absolute value.

Time of Testing versus Time of Teaching

In order to interpret test results in a useful and accurate manner, the interpreter first needs to be thoroughly aware of the content of the test in contrast to the scope and sequence of the instructional program being evaluated by the test. Test results should be interpreted in light of knowledge of when a concept is tested and when it is or will be taught. Skills taught after the testing program takes place obviously might be areas of poorer performance for some children. Weaknesses identified as such are still useful for instructional planning, but should not illicit alarm on the part of teachers, parents or students—instructional time is still available.

Mode of Testing versus Mode of Teaching

Remember that the measurement of a particular skill with a standardized achievement test is accomplished by sampling a child's behavior (performance) in just a few of the many ways in which that behavior could be observed (documented). As a result CTBS might measure some skills in a manner different from the manner emphasized during instruction of that same skill. Test results should be interpreted in light of these possible differences.

ctbs Comprehensive Tests of Basic Skills

Form U
Level C 42
Student Diagnostic Profile

DIRECTIONS

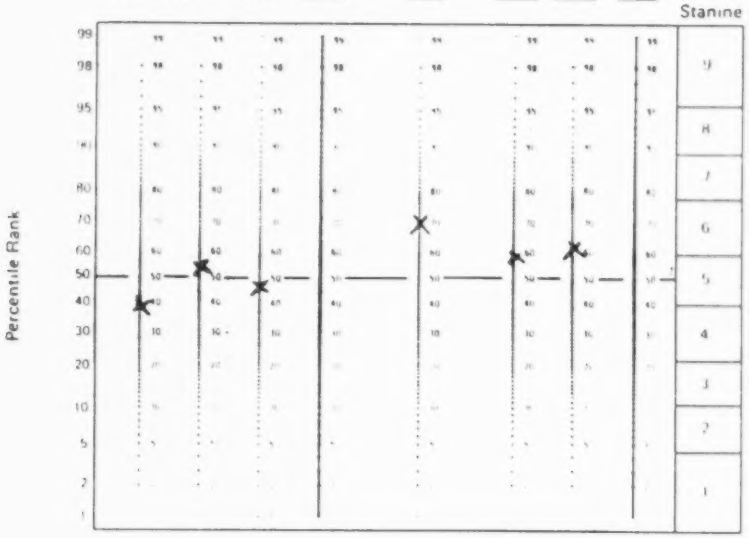
The CTBS U and V Student Diagnostic Profile provides a record of a student's test scores and is a source of information for instructional planning. Follow these instructions to complete Parts 1, 2, and 3 of the profile.

PART 1 STUDENT IDENTIFYING DATA: Record the student's name, teacher's name, grade, and test date in the appropriate spaces.

NAME JOHNNY
TEACHER MS. PROUD
GRADE 1
TEST DATE 5/11/86

PART 2 SUMMARY OF SCORES AND PERCENTILE RANK PROFILE:
(1) In the profile section below, enter the number of correct responses for each test in the spaces labeled NUMBER-CORRECT SCORE.
(2) Refer to the table in the appropriate Norms Book to convert the number-correct scores to scale scores. Scale scores for Total Content Areas are obtained by averaging the scale scores of the tests they include.
(3) Refer to the tables in the appropriate Norms Book to obtain other scores such as percentile ranks and grade equivalents. Enter these scores in the appropriate spaces. The corresponding stanine for a percentile rank may be determined by referring to the far right-hand column of the graph. Objectives mastery scores may be summarized from the information in Part 3.
(4) On the graph, mark a short, heavy line across the vertical bar at the point that corresponds to the percentile rank for each test and total. The position of these lines on the profile gives a graphic representation of the student's relative achievement in the test content areas.

Test	1 Word Attack	2 Vocabulary	3 Reading Comp	Total Reading (Tests 2 + 3)	4 Language Expression	5 Mathematics Compu	6 Mathematics Concepts and Applications	Total Mathematics (Tests 5 + 6)
Possible Number-Correct Score	30	25	25		20	20	25	
Number-Correct Score	24	20	18		18	18	20	
Scale Score*	493	493	456	475	502	466	529	498
Percentile Rank	39	53	48	50	70	59	60	61



Grade Equivalent		1.8	1.7	1.8	2.1	1.4	1.9	1.9
Normal Curve Equivalent	1.4	52	49	50	61	53	56	55
Objectives Mastery Score	/5	/4	/3	/7	/5	/2	/3	/5
Additional Score								

*Scale Score for Total Reading = Scale Scores for (Test 2 + Test 3) ÷ 2
Scale Score for Total Mathematics = Scale Scores for (Test 5+ Test 6) ÷ 2

Published by CTB/McGraw-Hill, Del Monte Research Park, Monterey, California 93940 Copyright © 1981 by McGraw-Hill, Inc. All Rights Reserved
Printed in the United States of America.

94922

Test on Test Construction

I. Directions: From the five answers (A, B, C, D, E) listed below, check the statement which most clearly pertains to the 18 statements and mark the chosen response on the answer sheet opposite the appropriate item number.

To which of the following types of test items do the statements listed below most clearly pertain?

A. Matching
B. Completion
C. Multiple Choice
D. True-False
E. Essay

1. Measures organizational ability of students
2. Is weaker in scoring reliability
3. Measures student's ability to recall facts
4. Is best for measuring association among terms and their definitions
5. Is best for quick testing of a number of associations
6. Is the most subjective of the objective types of questions
7. Is the best for measuring discrimination and understanding
8. Is the best for measuring the student's powers of synthesis
9. Is the hardest type of item to grade
10. Is the most time consuming in grading
11. A test composed of these items can be prepared quickly
12. Permits wide sampling or coverage of materials to be tested
13. Is susceptible to lifting statements out of context in constructing
14. Allows the same alternatives to be used more than once
15. Is the most difficult type of item to construct
16. Is the most susceptible to guessing by students
17. Is the most susceptible to the halo effect
18. Is least helpful in diagnosing pupil difficulties

II. In the following multiple choice items, select which one of the given responses best completes the statement or answers the question. Mark the appropriate corresponding blank on the answer sheet.

19. Which of the following is the best true-false item?
    A. Hamilton was a Federalist and a strict constitutionalist.
    B. Shakespeare did not write Hamlet.
    C. Always end a sentence with a period.
    D. In the early schools of America, examinations were given orally.

20. Which of the following is the best multiple choice item?
    A. The capital of the United States of America is (1) Boston (2) Washington, D.C. (3) Chicago (4) New York.
    B. Examples of fungi are (1) mushrooms (2) ferns (3) liverwort (4) gleocapsa.
    C. The horticulturist would classify the lady's slipper as an (1) carnation (2) gardenia (3) orchid (4) tulip.
    D. The greatest single contributor to college success is (1) intelligence (2) motivation (3) experience (4) health

21. Matching questions can be improved by
    A. Reading directions orally, rather than writing them on the test.
    B. Testing only one idea or relationship per unit or question.
    C. Including an even number of items in both columns.
    D. Lengthening the question to include more than 8 items.

22. Of the following, the least appropriate use of the multiple choice test is in connection with the measurement of the
    A. knowledge of basic facts.
    B. ability to apply knowledge.
    C. understanding of principles.
    D. its adaptability to the measurement of discrimination.

23. An important advantage of the multiple choice type test is its
    A. ease of construction.
    B. requirement of organization by pupils.
    C. emphasis upon recall rather than recognition.
    D. its adaptability to the measurement of discrimination.

24. The chief objection to the simple recall test is that
    A. it is somewhat lacking in objectivity.
    B. its use is restricted to the testing of specific facts.
    C. it is impossible to machine score.
    D. it does not permit easy handling of the guessing problem.

25. The chief "selling point" of an essay test is its
    A. ability to measure skill in organizing material.
    B. total economy of the teacher's time.
    C. high validity.
    D. high reliability.

BIBLIOGRAPHY

Ahmann, J. S., & Glock, M. D. (1981). Evaluating student progress: Principles of tests and measurements, 6th ed. Boston: Allyn & Bacon.

Kubiszyn, T., & Borich, G. (1984). Educational testing and measurements, classroom application and practice. Glenview, IL: Scott, Foresman & Co.

Popham, W. J. (1978). Criterion-referenced measurement. Englewod Cliffs, NJ: Prentice-Hall.